PATENT ABSTRACTS OF JAPAN

(11)Publication number : 03-116100
(43)Date of publication of application : 17.05.1991

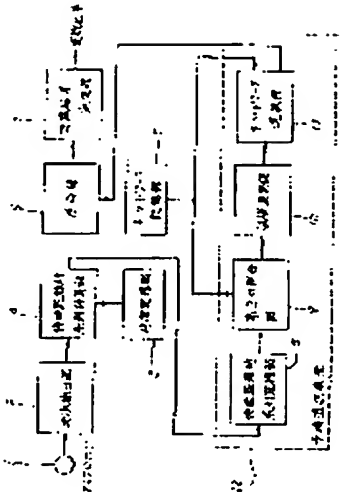
(51)Int.Cl. G10L 3/00

(21)Application number : 01-251812
(22)Date of filing : 29.09.1989

(71)Applicant : FUJITSU LTD
(72)Inventor : KIMURA AKIHIRO

(54) LARGE VOCABULARY VOICE RECOGNIZING DEVICE

(57)Abstract:
PURPOSE: To simply execute the real time recognition even in the case of a large vocabulary by providing a pre-selecting device, and compressing a feature distance time series being a series of a distance of each frame of an input voice and each fundamental unit.
CONSTITUTION: In a pre-selecting device 12, a feature distance time series compressing part 8 compress a feature distance time series calculated by a feature distance time series calculating part 4. A collating part 9 of a second part collates a network read out of a network storage part 5 and the compressed feature time series, and calculates a distance of both of them. A candidate selecting part 10 outputs vocabulary names of the network of the number of pieces determined in advance as a result of pre-selection in order of a small distance to inputs of each network obtained by the collating part 9. A network selecting part 11 reads out only the network of the vocabulary name obtained by the candidate selecting part from the network storage part 5, and transfers it to a collating part 6. In such a way, the voice recognition of a large vocabulary can be executed quickly.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

⑲日本国特許庁（JP）　⑪特許出願公開

⑫公開特許公報（A）　平3－116100

⑤1Int. Cl.⁵	識別記号	庁内整理番号	⑬公開　平成3年(1991)5月17日
G 10 L 3/00	301 D	8842－5D	

審査請求　未請求　請求項の数 6　（全8頁）

⑤4発明の名称　大語彙音声認識装置

⑳特　願　平1－251812

㉒出　願　平1(1989)9月29日

⑦2発明者　木村　晋太　神奈川県川崎市中原区上小田中1015番地　富士通株式会社内

⑦1出願人　富士通株式会社　神奈川県川崎市中原区上小田中1015番地

⑦4代理人　弁理士　本間　崇

明　細　書

1．発明の名称

大語彙音声認識装置

2．特許請求の範囲

1．入力音声区間の一定微小時間毎の特徴時系列を得る特徴抽出部(2)と、音声の各基本単位の特徴を記憶した特徴記憶部(3)と、各基本単位の特徴と入力音声の特徴時系列の距離を計算することにより各基本単位の特徴距離時系列を得る特徴距離時系列計算部(4)と、音節、単語、文節、または文章などの認識対象のテンプレートとして音声の基本単位のネットワークを予め記憶するネットワーク記憶部(5)と、入力音声区間の特徴距離時系列を予め用意した複数のネットワークと照合することにより各ネットワークと入力音声の距離を計算する照合部(6)と、計算された距離の最も小さいネットワークに対応する単語名等を認識結果として出力する認識結果決定部(7)を有する音声認識装置において、

前記、特徴距離時系列計算部(4)の出力である入力音声区間の一定微小時間毎の特徴距離時系列を圧縮する特徴距離時系列圧縮部(8)と、入力音声の圧縮された特徴距離時系列を予めネットワーク記憶部(5)に記憶されたネットワークと照合することにより各ネットワークと入力音声の概略距離を計算する第2の照合部(9)と、計算された概略距離の小さいものから予め決められた個数の単語等を選び出す候補選択部00と、候補選択部の結果に存在するネットワークのみをネットワーク記憶部(5)から読み出し照合部(6)に転送するネットワーク選択部00から成る予備選択装置を設けたことを特徴とする大語彙音声認識装置。

2．特徴距離時系列圧縮部は、特徴距離時系列の一定時間ごとの区間内の予め決められた位置から系列要素を標本化し、その標本値に区間を代表させることにより、特徴距離時系列

を圧縮する構成である請求項1記載の大語彙音声認識装置。

3．特徴距離時系列圧縮部は、特徴距離時系列の一定時間ごとの区間内の各音声単位の特徴距離の平均値を計算し、その平均値に区間を代表させることにより、特徴距離時系列を圧縮する構成である請求項1記載の大語彙音声認識装置。

4．特徴距離時系列圧縮部は、特徴距離時系列の一定時間ごとの区間内の各音声単位の特徴距離毎の最小距離を求め、その最小距離に区間を代表させることにより、特徴距離時系列を圧縮する構成である請求項1記載の大語彙音声認識装置。

5．入力音声の圧縮された特徴距離時系列と既知のネットワークとの照合に際して距離の代りに類似度を用い、類似度の大なるものを候補として選出する請求項1～4記載の大語彙音声認識装置。

6．入力音声の圧縮された特徴距離時系列と既知のネットワークとの照合に際して距離の代りに確率を用い、確率の大なるものを候補として選出する請求項1～4記載の大語彙音声認識装置。

3．発明の詳細な説明

［概　要］

大語彙の音声を認識する装置であって、入力音声に対する候補単語を高速に選び出す予備選択装置を有する音声認識装置に関し、

入力音声の各フレームと音声の各基本単位（子音、母音など）との距離の系列である特徴距離時系列の圧縮を行うことにより、認識処理量を大幅に削減する予備選択方式において、高い予備選択性能を得ることを目的とし、

特徴距離時系列の圧縮方式として、

「圧縮対象区間の予め決められた点で区間を代表させることにより圧縮を行う」か「圧縮区間の平均値で区間を代表させることにより圧縮を行う」か、または「圧縮区間の各基本単位の最小距離で区間を代表させることにより圧縮を行う」手段を設けることにより構成する。

［産業上の利用分野］

本発明は音声認識、特に非常に多くの認識対象を必要とする音声文書作成、あるいは、音声による商品名入力等に用いられる大語彙音声認識装置に関し、特に、照合に際する処理量を減少せしめて処理の高速化を図るための予備選択方式に係る。

［従来の技術］

第6図は、従来の音声認識装置の構成の例を示す図である。

同図において、51はマイクロホン、52は特徴抽出部、53は特徴記憶部、54は特徴距離時系列計算部、55はネットワーク記憶部、56は照合部、57は認識結果決定部を表わしている。

以下、各部の動作等について説明する。

マイクロホン1は入力された音響音声信号を電気音声信号に変換する。

特徴抽出部52は電気音声信号をデジタル化するとともに、電気音声信号を数ミリ秒～十数ミリ秒の間隔でＦＦＴ（高速フーリェ変換）などを用いて周波数分析する。

特徴記憶部53には音声の基本単位である各母音や各子音を予め分析した特徴を格納してある。

特徴距離時系列計算部54は特徴抽出部52で計算された分析結果と特徴記憶部53から読み出した各母音及び各子音の特徴との距離計算を行い、第7図で示されるようなフレーム58を生成し、入力音声の全体にわたって第8図に示されるようなしフレームから構成される特徴距離時系列（フレーム列）を生成する。同図において、59－1～59－Ｌはそれぞれフレームを表わしており、Ｌは発声長に相当する。

ネットワーク記憶部55には第9図に示されるようなネットワークが記憶されている。ネットワークは各単語の可能な複数種類の発音を表わしたものであり、単語の端を表す2個の一つの

パスが一種類の発音に対応している。第9図のネットワークは「愛知（アイチ）」という単語のネットワークであり、aが母音の「ア」、iが母音の「イ」、qが「チ」の前の閉鎖、chが「チ」の子音部分、その後ろのiが「チ」の母音部分、またciは無声化した（母音部分が発声されない）「チ」を表わしている。

照合部56はネットワーク記憶部55に記憶されている各語彙のネットワークと特徴距離計算部で得られた特徴距離時系列の照合を行い、各ネットワークと特徴距離時系列の距離を計算する。この照合は動的計画法（ＤＰ）を用いて行われる。照合部56は各ネットワークごとに入力（特徴距離時系列）との距離を計算し出力する。

認識結果決定部57は照合部56で得られた各ネットワークの入力との距離を小さい順にソーティングし、距離の小さい順にネットワークの語彙名を認識結果として出力する。

［発明が解決しようとする課題］

上述したような従来の方式においては、ネットワーク記憶部に記憶されているネットワーク数が数百程度までは実時間認識を行うことが可能であるが、それを越えると実時間認識ができなくなり、大語彙（数万～十万語）を認識する場合は実用上の問題点があった。

すなわち、この方式では、特徴距離系列計算部54は、特徴抽出部52が、入力された電気音声信号を数ミリ秒～十数ミリ秒の間隔で周波数分析して出力する全部のデータについて、これと特徴記憶部53に記憶されている各母音や子音の特徴との距離計算を行なった結果の膨大なデータを出力し、照合部56は、これとネットワーク記憶部55に記憶されているネットワークデータとを照合するので、その処理に多大の時間を必要とするのである。

そのため、従来から予備選択方式を導入して、この問題点を解決する方式が提案されているが、予備選択用の特別の辞書を用意する必要とする場合が多く、特に大語彙の場合にはあらたに予備選択用の辞書を用意するのは非常に難しい。

本発明はこのような従来の問題点を鑑み、第6図に示したような従来の音声認識装置の構成を改良し、特別な予備選択用の辞書を必要としない予備選択方式を実現することにより、大語彙の場合にも簡単に実時間認識を行なうことのできる音声認識装置を提供することを目的としている。

［課題を解決するための手段］

本発明によれば、上述の目的は、前記特許請求の範囲に記載された手段により達成される。すなわち、本発明は、入力音声区間の一定微小時間毎の特徴時系列を得る特徴抽出部と、音声の各基本単位の特徴を記憶した特徴記憶部と、各基本単位の特徴と入力音声の特徴時系列の距離を計算することにより各基本単位の特徴距離時系列を得る特徴距離時系列計算部と、音節、単語、文節、または文章などの認識対象のテンプレートとして音声の基本単位のネットワークを予め記憶するネットワーク記憶部と、入力音声区間の特徴距離時系列を予め用意した複数のネットワークと照合することにより各ネットワークと入力音声の距離を計算する照合部と、計算された距離の最も小さいネットワークに対応する単語名等を認識結果として出力する認識結果決定部を有する音声認識装置において、前記特徴距離時系列計算部の出力である入力音声区間の一定微小時間毎の特徴距離時系列を圧縮する特徴距離時系列圧縮部と、入力音声の圧縮された特徴距離時系列を予めネットワーク記憶部に記憶されたネットワークと照合することにより各ネットワークと入力音声の概略距離を計算する第2の照合部と、計算された概略距離の小さいものから予め決められた個数の単語等を選び出す候補選択部と、候補選択部の結果に存在するネットワークのみをネットワーク記憶部から読み出し照合部に転送するネットワーク選択部を有する予備選択装置を設けた音声認識装置

である。

［作　用］

第1図は本発明の原理的構成を示す図であって、1はマイクロホン、2は特徴抽出部、3は特徴記憶部、4は特徴距離時系列計算部、5はネットワーク記憶部、6は照合部、7は認識結果決定部を表わしており、これらによって構成される音声認識の原理は第6図に基づいて説明した従来のものと概ね同様である。一方、特徴距離時系列圧縮部8、第2の照合部9、候補選択部10、ネットワーク選択部11によって構成される点線で囲んだ部分が本発明の特徴を成す予備選択装置12を示している。

同図において、特徴距離時系列圧縮部8は特徴距離時系列計算部4で計算された特徴距離時系列を圧縮する。圧縮の様子を第2図に示す。同図において13は特徴距離時系列計算部4で計算された特徴距離時系列を示しており、Lフレームからなる。また14は4フレーム区間毎に圧縮された特徴時系列の例を示している。

第2の照合部9は、照合部6（第6図における照合部56と同様）と同じ動作を行う。すなわち、ネットワーク記憶部から読み出されたネットワークと圧縮された特徴時系列の照合を行い、各ネットワークと圧縮された特徴距離時系列の距離を計算する。この照合は例えば動的計画法（ＤＰ）を用いて行われる。第2の照合部9は各ネットワークと圧縮された特徴時系列との距離を計算し出力する。

候補選択部10は第2の照合部9で得られた各ネットワークの入力との距離を小さい順にソーティングし、距離の小さい順に予め決められた個数（例えば 500個）のネットワークの語彙名を予備選択結果として出力する。

ネットワーク選択部11は、候補選択部10で得られた語彙名のネットワークのみをネットワーク記憶部5より読み出し、照合部6に転送する。

このように構成することにより、例えばネットワーク記憶部に1万語のネットワークがある場合に従来の方法では2秒の処理時間（一般に処理時間が 0.3秒以下であれば実時間認識と呼ぶ）がかかるとすると、10フレームを1フレームに圧縮する特徴距離系列圧縮部を用いれば第2の照合部9の処理量が従来の照合部の1/10（0.2秒）となり、候補選択部で 500個の候補を出力することにすると、照合部6では、その 500個のネットワークの照合を行うだけでよいので照合処理時間は 500／10000 × 2秒＝ 0.1秒であり、合計 0.3秒の照合処理時間で認識を行うことができるから、実時間認識が容易に実現できることになる。

［実施例］

本発明による音声認識装置の原理的構成は第1図に示したとおりであり、各部が上述したような動作を行なうことにより特徴距離時系列の圧縮を行なって候補の数を削減して、大語彙音声の認識を高速で行なうものであるが、本発明においては、その特徴距離時系列の圧縮方法についての特徴がある。以下、これについて実施例に基づいて説明する。

第3図は第1の実施例を説明する図であって、(a)は特徴距離時系列圧縮部の構成の例を、(b)は標本化の例を示しており、15は区間バッファ、16は標本化部、17は特徴距離時系列計算部から出力された一区間の特徴距離時系列のフレーム群、18は圧縮された特徴距離時系列のフレームを表わしている。

本実施例は、圧縮区間内の予め決められた場所を標本化することにより圧縮を実現するもので、区間バッファ15は、圧縮すべき区間を一時的に記憶する。標本化部16は区間バッファ15の予め決められたアドレスの内容のみを読み出し出力する。本例においては(b)に示すように、一区間の特徴距離時系列のフレーム群17から、その先頭のフレームを圧縮データ18として抽出して、これによって一区間の特徴距離時系列17を代表するフレームとする場合を示している。

圧縮されたフレームとして抽出するのは先頭

フレームに限るものではなく、予め定めた任意の一定の位置のフレームでも良く、また、その都度何らかの要因に基づいて決定した任意のフレームであっても良い。

第4図は第2の実施例を説明する図であって、(a)は特徴距離時系列圧縮部の構成の例を、(b)は平均計算についての説明を示しており、19は区間バッファ、20は平均計算部、21は一区間の特徴距離時系列のフレーム群、22は上記一区間の特徴距離時系列のフレームの各値の平均値を採ることにより圧縮されたフレームを表わしている。すなわち、本実施例は、圧縮区間内の各音声単位の特徴距離の平均で区間を代表させることにより圧縮を実現するもので区間バッファ19に格納された一区間の各フレームについて平均計算部20で各音声単位の特徴距離毎に平均値を計算し出力することにより、これらの平均値を有する圧縮フレーム22を生成するものである。

第5図は第3の実施例を説明する図であって、(a)は特徴距離時系列圧縮部の構成の例を、(b)は最小距離検索による圧縮の例を示しており、23は区間バッファ、24は最小距離検索部、25は一区間の特徴距離時系列のフレーム群、26は圧縮されたフレームを表わしている。また、英字符A～Dで示す黒丸印は区間中で音声の各基本単位との最小の距離の値を持つフレーム位置を示している。

本実施例は、圧縮区間内の各音声単位の特徴距離の最小距離で区間を代表させることにより圧縮を実現するもので区間バッファ23に格納された一区間の各フレームから最小距離検出部24が各音声単位の特徴距離毎に最小距離を検出して出力することにより、各要素がそれぞれ最小距離を有する圧縮フレーム26を生成するものである。

なお、以上の説明では総て、未知の音声に係る特徴時系列と、既知の音声の各基本単位の特徴とを比較してその距離を求め、あるいはその距離の値の最も小なるものを採択するものとして説明しているが、これらは、距離に限るものではなく、両者間の類似度または確率を求め、その値の最大なるものを採択する方法を採る場合であっても全く同様な効果が得られることは明らかである。

［発明の効果］

以上説明したように本発明によれば、予備選択用の辞書等を用意することなく、簡潔な構成の音声認識装置によって大語彙の音声認識を迅速に行なうことができる利点がある。そして、データを圧縮したことによる認識率の低下も少なく、例えば、前述の第3の実施例の方法（特徴時系列の一区間内の各フレームの内の各音声単位の特徴距離ごとの最小距離を求める方法）を用いて、実験した結果の例では、1000単語（地名）を用い、男女各5名の話者で、音声の基本単位の特徴を学習するための学習単語数を200としたとき、特徴時系列圧縮部での圧縮率が30／1000（候補30個）の場合の誤り率が0.2%以下と言うデータが得られている。

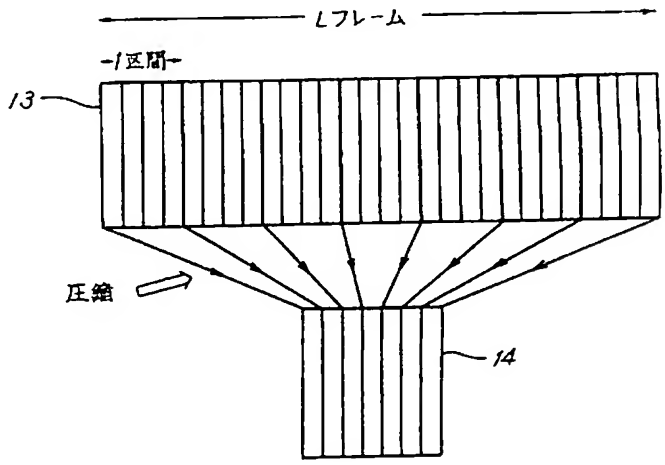
4．図面の簡単な説明

第1図は本発明の原理的構成を示す図、第2図は特徴時系列の圧縮について説明する図、第3図は第1の実施例を説明する図、第4図は第2の実施例を説明する図、第5図は第3の実施例を説明する図、第6図は従来の音声認識装置の構成の例を示す図、第7図はフレームの構成の例を示す図、第8図は特徴距離時系列（フレーム列）の例を示す図、第9図はネットワークの例を示す図である。

1……マイクロホン、2……特徴抽出部、3……特徴記憶部、4……特徴距離時系列計算部、5……ネットワーク記憶部、6……照合部、7……認識結果決定部、8……特徴距離時系列圧縮部、9……第2の照合部、10……候補選択部、11……ネットワーク選択部、12……予備選択装置、13……特徴距離時系列、14……圧縮された特徴距離時系列、15，19，23……区間バッファ、16……標本化部、17，21，25……1区間の特徴

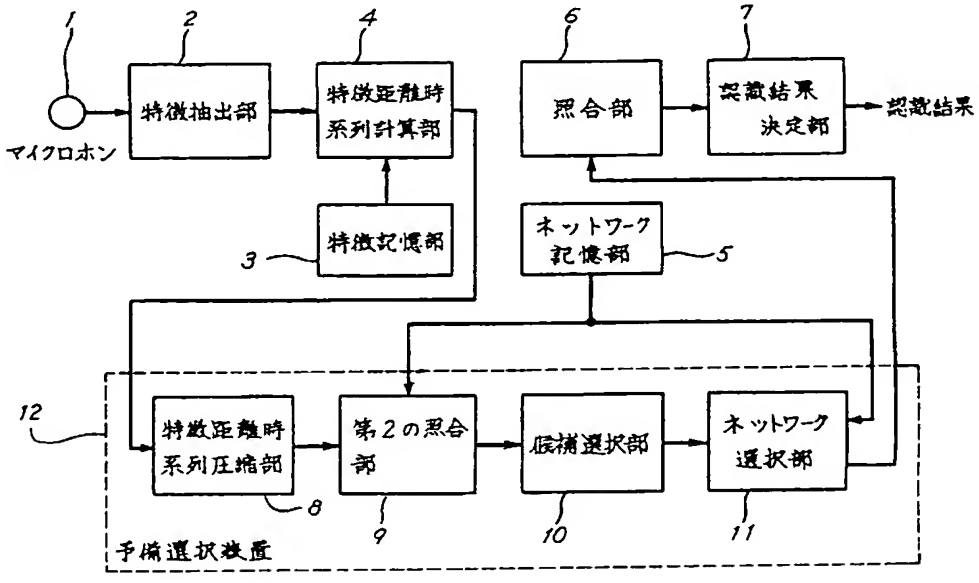
距離時系列データのフレーム群、18，22，26…
…圧縮された特徴距離時系列のフレーム、20…
…平均計算部、24……最小距離検索部

代理人　弁理士　　本間　　崇



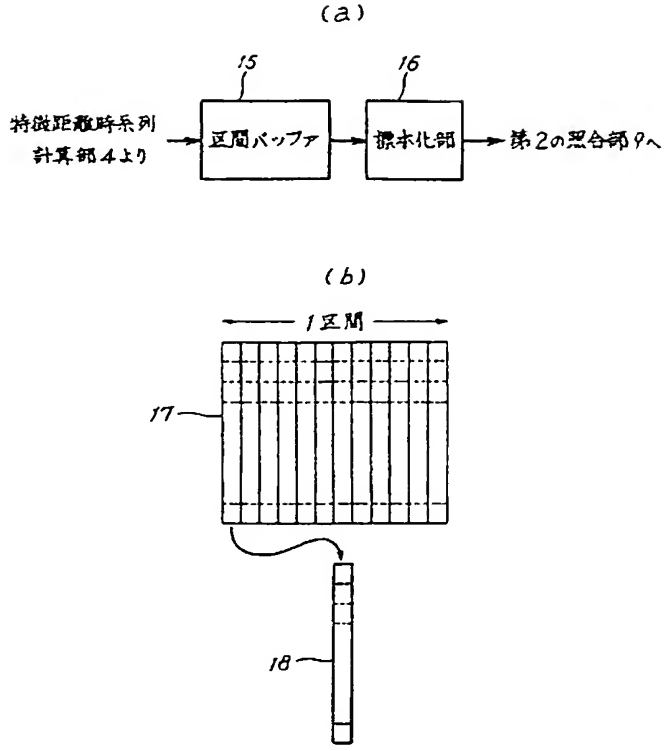
特徴距離時系列の圧縮について説明する図

第　2　図



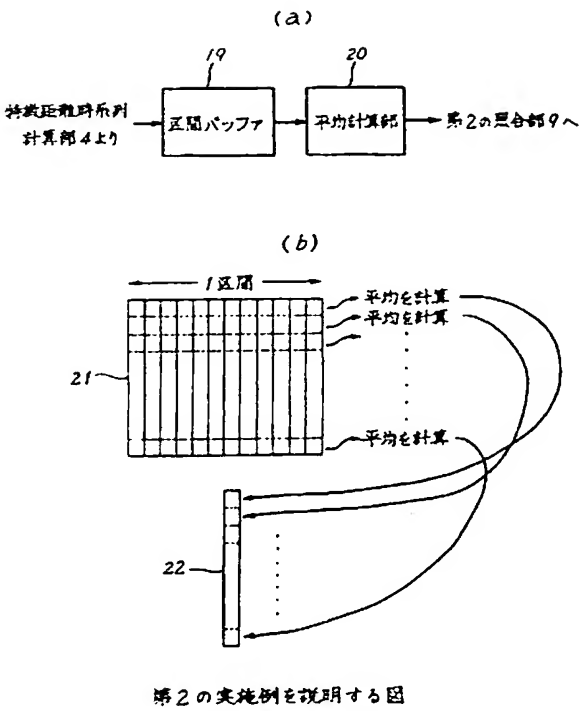
本発明の原理的構成を示す図

第　1　図



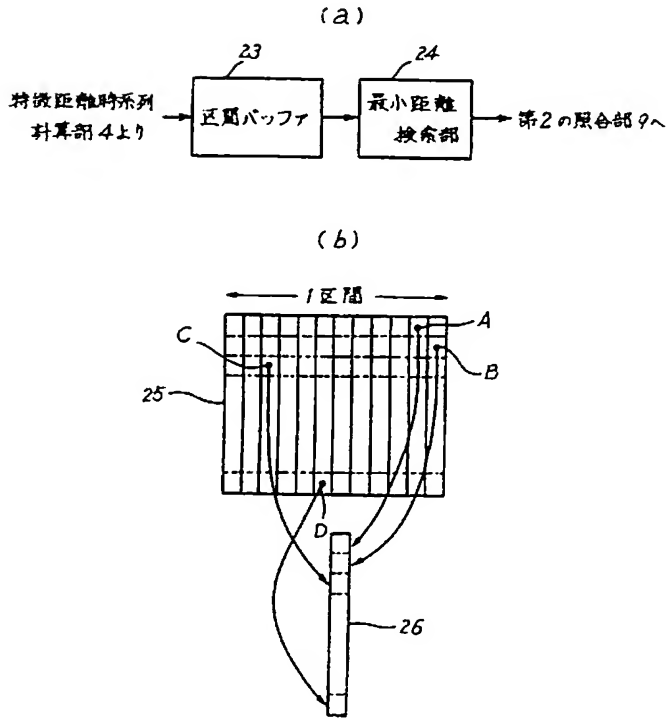
第1の実施例を説明する図

第 3 図



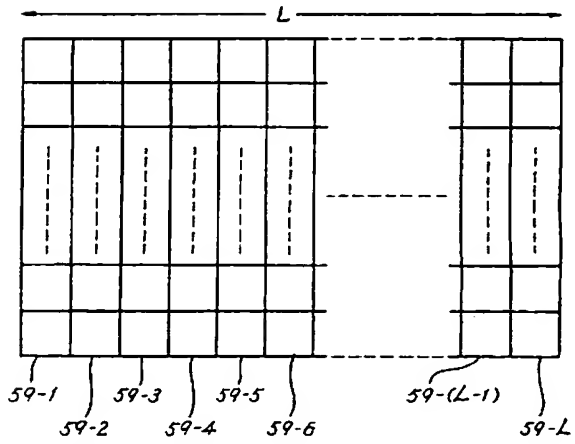
第2の実施例を説明する図

第 4 図



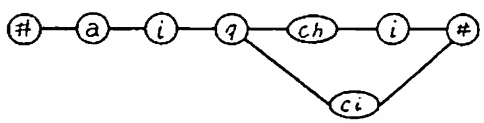
第3の実施例を説明する図

第 5 図



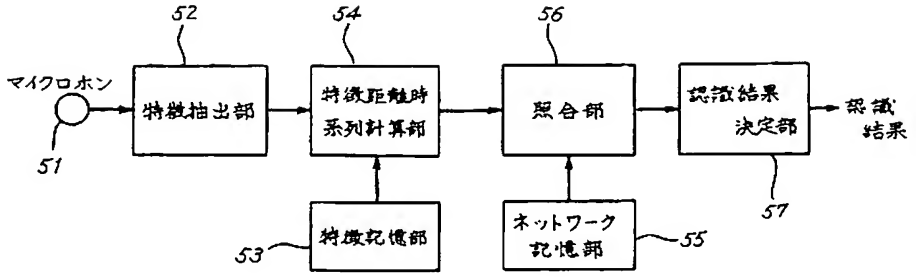
特徴距離時系列（フレーム列）の例を示す図

第 8 図



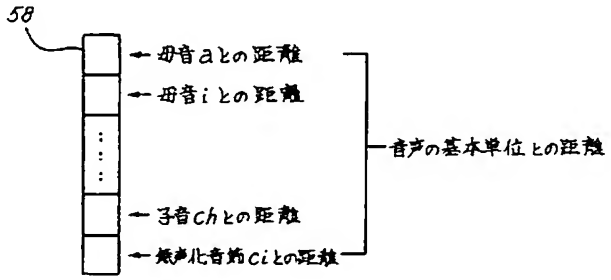
ネットワークの例を示す図

第 9 図



従来の音声認識装置の構成の例を示す図

第 6 図



フレームの構成の例を示す図

第 7 図